International Journal of Cognitive Computing in Engineering 6 (2025) 267–279

Contents lists available at ScienceDirect

International Journal of Cognitive Computing in Engineering

journal homepage: http://www.keaipublishing.com/en/journals/international-journal-of-cognitive-computing-in-engineering/

Prenatal depression level prediction using ensemble based deep learning model

Abinaya Gopalakrishnan [a,b,*], Xujuan Zhou [a], Revathi Venkataraman [b], Raj Gururajan [a,b], Ka Ching Chan [a], Guohun Zhu [a], Niall Higgins [c]

[a] School of Business, University of Southern Queensland, Springfield, 4300, Queensland, Australia
[b] Department of Networking and Communications, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, 603203, TamilNadu, India
[c] Royal Brisbane and Women's Hospital, Herston, 4006, Queensland, Australia

ARTICLE INFO

Keywords:
Electrodermal activity (EDA)
Childbirth stress
Stress detection
Wearable device

ABSTRACT

Background and objective: Many people find that the emotional and mental strain of labor and delivery is greater than they anticipated. However, there are few reports on stress levels during pregnancy, and there is limited research into stress observation during delivery. Prenatal depression during the delivery has to be monitored continuously without disturbing the mothers during the childbirth.

Methods: We explore the potential of employing EDA for Prenatal Depression prediction. The proposed model applies a novel method for motion artifacts followed by data labeling using PHQ-9 score values and LOOCV applied to train robustly. This culminated in the development of a novel EBDL model to accurately predict stress levels.

Results: We subsequently applied the ensemble based deep learning model on a testing dataset and our method proved to be 93.87 percent accurate, proving its superiority over the standard supervised classification models. The accuracy of this approach applied to three benchmark datasets produced better results compared to all commonly applied machine learning models, including an Ensemble based Deep Learning model.

Conclusion: The preliminary results are promising, and indicate a superior utility of EDA for monitoring stress levels in real-life scenarios. This approach should be applied to a clinical setting, it potentially could continuously monitor stress levels in pregnant women and provide real-time feedback of clinically important data for clinicians.

1. Introduction

Every person goes through stressful times, and those experiences help them become more equipped to handle future challenges. As a time of profound transformation, pregnancy places the mother under considerable strain, whether she is aware of it or not. However, persistent stress has been associated with unfavorable health implications such as depression which is detrimental to the health of the mother, infant and can even have a significant impact on family dynamics (Gopalakrishnan et al., 2023; Roberts et al., 2006). Some research indicates that depression during pregnancy can cause premature labor and postpartum depression can slow baby development (Gopalakrishnan, Venkataraman, Gururajan, Zhou and Zhu, 2022; Grace, Evindar, & Stewart, 2003; McMahon, Barnett, Kowalenko, & Tennant, 2006). Uterine contractions, cervical dilation, and effacement cause some of the most excruciating pain a woman will ever feel. Knowing what to expect during labor and delivery can help alleviate anxiety and boost confidence, all of which are important for a smooth delivery. Anger during labor increases awareness of pain, lengthens the process, and triggers the release of stress hormones that block oxygen supply to the uterus. Subsequently, uterine contractions can slow down, resulting in extended labor. The severity of labor pain varies from woman to woman and the degree of discomfort can range from mild to severe, corresponding to the mother's level of anxiety.

Recent advancements in measuring techniques, including hardware and software technologies, have pave the way for remote patient monitoring (Gopalakrishnan, Venkataraman, Gururajan, Zhou and Genrich, 2022). An individual's emotional and physical health, interpersonal dynamics, and general happiness have been studied using wireless sensing technologies. Various mobile applications have also been utilized for self-response psychological questionnaires. However, it is crucial to

* Corresponding author at: School of Business, University of Southern Queensland, Springfield, 4300, Queensland, Australia.
E-mail address: abinayag2@srmist.edu.in (A. Gopalakrishnan).

https://doi.org/10.1016/j.ijcce.2024.12.002
Received 23 January 2024; Received in revised form 11 December 2024; Accepted 15 December 2024
Available online 23 January 2025

remember that, while some stress cannot be avoided in our daily lives, untreated stress can have harmful consequences. Understanding the early stages of increased stress levels is an effective stress management method. Early detection allows mitigation steps to be implemented, reducing severity and making it easier to manage.

Reducing the intensity of the stress reaction produced by delivery has the potential to improve post-delivery health states, shorten hospital stays, in relation to labor length, pain levels, and improve patient care (Hobel, Goldstein, & Barrett, 2008). Women's mental health can be improved if they are exposed to less anxiety during the labor and delivery process (Federenko & Wadhwa, 2004). Therefore, less obstetric difficulties, longer gestation, a drop in the number of cesarean sections, and fewer post-partum complications, all of which contributed to an improvement in pregnancy outcomes. Masoudi, Kasraeian, and Akbarzadeh (2022) found that there was a drop in both the fetal heart rate and the fetal level of motor activity after relaxation, which was considered a good result.

Perinatal depression is commonly measured with one of three instruments: the Postpartum Depression Screening Scale (PDSS), the Patient Health Questionnaire PHQ-9 (Gopalakrishnan, Venkataraman, Gururajan, Zhou, Zhu, 2022), the Beck Anxiety Inventory (BAI). Most stress questionnaires are self-administered and have been shown to be accurate, but still require the respondent's full attention and participation. In addition, these options do not provide real-time feedback or the ability to monitor continuously. So, it is worthwhile exploring other alternative disruptive technologies to achieve the same goals. Electrodermal activity (EDA) is an all-encompassing term used to indicate the electrical properties of the skin. Unlike other organs of the human body, which are connected to the sympathetic and parasympathetic nervous systems, the skin with sweat glands and blood vessels is exclusively innervated by the sympathetic branch [9]. This makes EDA an ideal and unperturbed measure of sympathetic activation and therefore, the stress response, compared to other physiological measures like heart rate variability or blood pressure. Hence, this study focuses on an EDA based scheme for stress detection.

Whilst it is important to measure depression during pregnancy, as it can predict the level of depression postpartum (Beck, 1998; Nielsen, Videbech, Hedegaard, Dalby, Secher, 2000) there is some research that depression can also be exacerbated with high levels of stress (Sandman, Glynn, & Davis, 2016). Measurement and assessment of the level of depression during pregnancy is relatively straightforward using survey instruments such as PHQ-9; however, traditionally stress levels are measured using invasive techniques to assess levels of cortisol from urine, saliva, blood, or amniotic fluid samples (Caparros-Gonzalez et al., 2017). There is a clinical need to monitor physiological and psychological stress responses to labor not only as a reference to evaluate complicated pregnancies (Miller et al., 2019), but also because this can also predict future postpartum depression (Ayers, 2004). However, using cortisol as a measure of stress during labor may not be the best measure of psychological stress due to confounding natural physiological increases in cortisol. When a person experiences stress, the hypothalamic-pituitary-adrenal (HPA) axis is activated, which in turn triggers the creation of cortisol in the body. This is viewed as protective of the mother and child and promotes normal labor progression (Benfield, Newton, Tanner, & Heitkemper, 2014). What is warranted is an alternative non-invasive technique to continuously monitor stress levels during labor as a method to predict the future onset of postpartum depression.

The objective of this study is to predict the severity of depression in prenatal women using the ensemble-based deep learning model. The contributions of this study can be classified into four main entities as follows: (1) A wrist-wearable Ensemble Based Deep Learning (EBDL) model for stress level prediction. (2) A noise reduction technique that achieves 97. 83% accuracy, which employs an autoregressive model to identify motion artifacts in wrist EDA signal data. (3) A Leave One Out of Cross-Validation (LOOCV) method used during the training phase on a partitioned dataset to create a hybrid subject-dependent and subject-independent model with EDA to address individual differences sweat gland density and skin thickness. (4) Comparison of the accuracy of the cleaned EDA signal collected dataset with various EDA signals from the benchmark datasets such as CLAS, VerBIO, and WESAD datasets using the traditional as well as the EBDL model

The following sections of this paper are structured as follows: the related work will be briefly discussed in Section 2. In Section 3, the collected datasets that were used in the trials are discussed. These datasets include the PHQ-9 questionnaire that was given to mothers with symptoms of depression. Section 4 describes the proposed Ensemble Based Deep Learning (EBDL) model with a motion artifact detection architecture. The findings and a discussion of the suggested architecture for the assessment of prenatal depression are presented in Section 5. In Section 6, the conclusion is presented.

2. Related works

There are two main types of research that have been carried out to quantify stress in the field of human studies: laboratory studies and outdoor studies. The term "lab experiments" is commonly used to describe conventional stress studies in a laboratory setting. Stress tests carried out in a natural setting are known as "field experiments" (Salkind, 2010). The most important field studies published in the last ten years are summarized in Table 1. Research has shown that the majority of efforts to detect stress in daily life or at the workplace have relied on numerous physiological and behavioral data collected from participants in their everyday surroundings. However, stress monitoring based on physiological or behavioral abnormalities in a clinical population has not yet been investigated. Most of these studies used wearable technology to covertly record a variety of physiological signs in real time. Such datasets are always dense with information and may yield substantial understandings of stress's effects on everyday life and the workplace. However, the signals from these wearable sensors are susceptible to electrical noise and aberrations, which could impact the validity of any subsequent data analysis. As can be seen in the Table 1, most studies did not elaborate on how they handled motion.

Any wearable on the wrist carries the risk of introducing motion artifacts (MA) into EDA data for a variety of reasons. These include alterations in wrist rotation or hand movement, modifications in the tightness of the wearable device on the skin, and fluctuations in the amount of pressure given to the EDA sensors. Filtering, exponential smoothing, and wavelet-based adaptive de-noising have all been investigated by researchers as potential methods to remove artifacts (Chen et al., 2015; Hernandez, Riobo, Rozga, Abowd, & Picard, 2014; Poh, Swenson, & Picard, 2010). One major limitation of MA suppression methods is that they corrupt the entire time series, not just parts with artifacts, because they filter the data without any specificity. To address this, the MA detection approach was developed. Using machine learning, this method attempts to precisely encode the domain experts' understanding of MA detection within a classifier model.

Many MA detection algorithms (Taylor et al., 2015; Xia, Jaques, Taylor, Fedor, & Picard, 2015; Zhang, Haghdan, & Xu, 2017) have been published in the literature as supervised, semi-supervised, or unsupervised. When using only EDA attributes as input to the classifier, Taylor et al. (2015) supervised MA detection strategy based on SVM achieved an accuracy of 95 7%. The accuracy of an unsupervised technique using a k-NN clustering heuristic on the characteristics of the EDA was reported to be 89.8 percent by Zhang et al. who compared supervised and unsupervised MA identification systems (Zhang et al., 2017). Xia et al. show that semi-supervised learning processes can outperform supervised learning algorithms, which is a huge improvement over manually discovering the MA time periods in the EDA data (Xia et al., 2015). The supervised approach yielded an accuracy of 95.2%, while their test accuracy obtained was 95.8% is marginally higher.

Table 1
Performance comparison of stress related studies.

Ref No	Observing factor	Analysis (Best result**)	Highlights
Saylam and Incel (2024)	Daily life	Multivariate analysis with logistic regression	Multitask learning relieves depression and stress but not anxiety with RF and XGBoost.
Ma, Liu, and Yang (2024)	Work stress	Multimodal fusion model based on EEG, ECG and EDA reached 85.72%	Biased data collection; study cannot be universally applicable
Rykov et al. (2024)	Daily life stress	Data from a wearable HR and EDA utilized to determine the effects of stress on the user's uses a random differential GWO algorithm for feature extraction and ML algorithm called RF 95.6 percent.	User specific models are into consideration
Oubrahim, Amirat, Benbouzid, and Ouassaid (2023)	Daily life stress	EMAs and passive mobile logging is used for prediction; Binary classification; F1 score of 70% with RF classifier using PPG & contextual data	No MA removal; ISV is not considered
Zhu et al. (2023)	Daily life stress	Data from EDA systems are gathered, and SVM is found to be the most effective machine learning technique for predicting stress, with an accuracy of 92.9%.	Manual MA removal; ISV is not considered
Naegelin et al. (2023)	Work stress	The 10-fold cross-validation of the characteristics extracted from the mouse, keyboard, and HRV results in an average F1 rating of 0.775 for recognizing the three phases of emotional stress, valence, and alertness using support vector machines, random forests, and gradient boosting models.	No MA removal; ISV is not considered

The most helpful physiological indicator for the identification of stress and anxiety is the variation of heart rate (HRV) (Hong et al., 2010). Most existing devices measure stress using average Heart rate (HR), which is not as precise as Heart Rate Variability (HRV) parameters but still useful. While adjunctive Electroencephalogram (EEG) improves stress detection accuracy (Ahn, Ku, & Kim, 2019), it will be important for future research to determine whether dual technologies are useful for monitoring chronic stress over the long term. Kim, Park, and Park (2020) found that ElectroDermal Activity (EDA) was the best portable measure for stress detection because of its easy way to set up and use; EDA is a generalized information for the electrical characteristics of the skin. Unlike other organs in the human body that are connected to the parasympathetic and sympathetic nervous systems, the sympathetic branch of the nervous system totally innervates the skin, which contains sweat glands and blood arteries (Roy, Boucsein, Fowles, & Gruzelier, 2012). As a result of being an ideal and unaltered measure of sympathetic activation and the depression response, EDA stands out among other psychological metrics like variability in heart rate or blood pressure. Consequently, the emphasis of this research is placed on an EDA-based technique for the detection of depression. Therefore, electrodermal activity, also known as EDA, was recorded from the wrist of mothers to predict an automatic measurement of depression experienced during delivery. When it comes to detecting Alzheimer's disease, the deep model is more accurate and efficient (Gao & Lima, 2022). An accuracy of 64.15% was established by the generalizability of existing emotion detection systems that employ electroencephalography. These approaches were able to recognize positive, neutral and negative emotional states in individuals (Huangfu & Cheng, 2025).

2.1. Summary of research gaps

Recording electrodermal activity (EDA) is a strong and commonly used approach to study arousal in the mind or body. However, EDA's sensitivity to motion artifacts makes analysis difficult. When using physiological data to train machine learning algorithms, the high degree of inter-subject heterogeneity in physiological signals is a substantial barrier. The majority of the research on multilevel stress classification is founded on a generic, one-size-fits-all paradigm that assumes that people's responses to varied degrees of stress are constant across the board. Table 1 shows that in order to overcome this problem, studies have focused on developing topic-specific models. Classification accuracy may be guaranteed by such models for one topic, but may not work for others. The strategy requires building many models for each subject, making it difficult to choose the most appropriate classification for new data sets. This study presents an Ensemble-Based Deep Learning (EBDL) model that is based on signals to handle both subject dependency and independency in EDA data caused by factors such as sweat gland density and skin thickness. The method was created to address a problem that had been identified in earlier studies.

3. System overview

An overview of the dataset utilized for the study's experiments is given in this section. Details the main features of the dataset, the task with which it is associated, and the evaluation criteria.

3.1. Dataset collection

All eligible study participants were those admitted to the hospital after experiencing pain prior to the onset of labor or an adverse event had occurred such as an amniotic sac rupture. All who had a delivery and were predicted to have a history of depression were invited to participate in the study. This time slot was chosen to reduce the potential for diurnal changes such as feeling more intense symptoms of melancholic depression in the morning and to notice a gradual improvement as the day goes on. Physiological data was collected from each individual patient upon arrival at the delivery ward and after receiving their first administration of dilation treatment, as shown in Fig. 1. The practical aspects of the implementation, such as integration with existing healthcare systems, were quite easy, since the data was collected from a mobile device, so it does not require any special consideration.

3.2. Ethical clearance

The data collection for this research was authorized by the Institutional Ethics Committee (IEC) of the SRM Medical College and Research Center (SRMC & RC), Chennai, India. Ethical standards and regulations were followed throughout all phases of data collection. By signing a consent form, each participant confirmed her understanding and acceptance of participating in the study. Data were collected in 2022, between April and December.

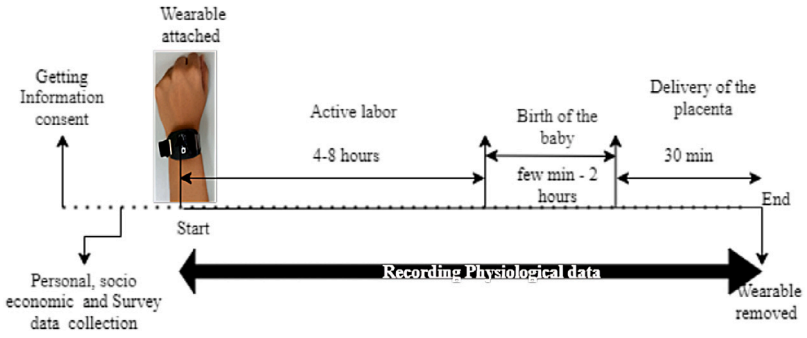

Fig. 1. Survey based personal, socioeconomic and depression data collection.

3.3. Participants selection

Women admitted to SRMC & RC were chosen for the study using a sequential participant selection technique. This made it possible to collect information from women at a crucial point in the delivery process. This also improved the chances of early detection of symptoms of elevated stress and provided effective treatment by clinicians.

3.3.1. Criteria for acceptance of subjects

Participants were informed about the purpose of the study and all met the following inclusion criteria before giving their informed consent:

- women who have given birth between the ages of 19 and 35.
- The patient's information and consent form were acknowledged by the participants as being understood.
- The number of pregnancies a mother has is not related to the method of birth (spontaneous vs. induced).

The objectives of the criteria were to ensure the best opportunity to identify the range of mothers with varying ages and to include mothers with a range of parity and delivery styles.

3.3.2. Exclusion criteria

The following conditions precluded mothers from taking part in the study:

- Mothers carrying more than one baby.
- Mothers who had undergone in vitro fertilization (IVF) treatment.
- Mothers with a history of complications or adverse events in obstetrics.
- Mothers whose pregnancies were deemed high-risk by clinicians were also not eligible to participate in the study. The present conditions included those with preeclampsia, chronic disease, gestational diabetes mellitus, or fetal abnormalities.

Unfortunately, this may have included those who had an increased known risk of postpartum depression or who had atypical responses to elevated stress. These mothers may also have had additional needs that could prevent them from dedicating the time and effort required to complete the investigation.

3.3.3. Patient Health Questionnaire PHQ-9

The Patient Health Questionnaire's PHQ-9 is a quick self-report test that incorporates a depression rating scale and DSM-IV depression diagnostic criteria. The patient can complete the PHQ-9 in under ten minutes, and the clinician can quickly grade it. It is also possible to administer it multiple times to capture fluctuations in the severity of depression as a result of treatment. To quantify the intensity of depression symptoms, a raw score is used, which can vary between 0 and 27 (He et al., 2020). The following is a description of the different severity levels: Mild depression — 5 to 9 Moderate depression — 10 to 14 Moderately severe depression — 15 to 19 Severe depression — 20 to 27

3.3.4. Cortisol in saliva as a biomarker

When a person experiences stress, the hypothalamic-pituitary-adrenal axis (HPA) is activated, which in turn triggers the creation of cortisol in the body. Salivary cortisol is widely accepted to be a biomarker for stimulation of the sympathetic system during times of stress (Anusha et al., 2019). To determine the levels of cortisol that were present in the systems of mothers before and after birth, several samples of their saliva were obtained throughout delivery.

3.4. EDA signal and characteristics

Electrodermal activity is a human property that causes the electrical characteristics of the skin to change constantly. Sweat secretion causes fluctuations in skin conductivity, which can be monitored with an EDA sensor. An EDA signal is characterized by its tonic and phasic values. Both the tonic and phasic levels have characteristics that are polar opposites of each other. The moisture of the skin of each person and the innate adaptability of the person affect the tonic level, which varies gradually and smoothly in the EDA. The tonic level, often known as the skin conductance level (SCL), is the first parameter measured in the analysis of electrical dynamics. In EDA, the phasic level is the component with the fastest reaction time and the most robust response to stimuli (Bogomolov, Lepri, Ferron, Pianesi, & Pentland, 2014). The skin conductance response (SCR) is a biphasic indication of the physiological state. Changes in skin conductance (SCR variations) are caused by perspiration that forms when a person experiences a problem. Phasic level shifts in SCR are typically more pronounced and rapid than tonic level fluctuations in SCL. Data show spikes or bursts indicative of SCR variations. Skin conductance responses can be event- or stimulus-specific. In most cases, ER-SCRs appear between 1 and 5 s after stimulation. However, NS-SCRs occur in the absence of any external trigger or awareness. The presence of NS-SCRs in the EDA raw signal makes direct quantification of the SCL challenging, regardless of whether people receive planned stimuli or not. Filters are used to separate the SCL and SCR components of the raw EDA signal during processing. SCR amplitudes and timings of onset and offset.

Fig. 2. A signal with EDA has two parts and four characteristics. (1) Latency: The time it takes for the stimulus to begin and for the phasic burst to begin. (2) Extreme magnitude: The magnitude of the transition from the beginning to the highest point. (3) Rise time: How long does it take before the peak occurs after the start? (4) Rest time: How long does it take to get back to peak performance?

3.5. Hardware description

The recording of physiological data was carried out with the help of a battery operated wrist wearable developed by Analog Devices (Broeders, 2017), hereafter referred to as Analog Device Vital Sign Monitoring (ADI-VSM). This wristwatch is capable of constantly recording electro-cardiogram (ECG) signals, skin temperature (ST), photoplethysmogram

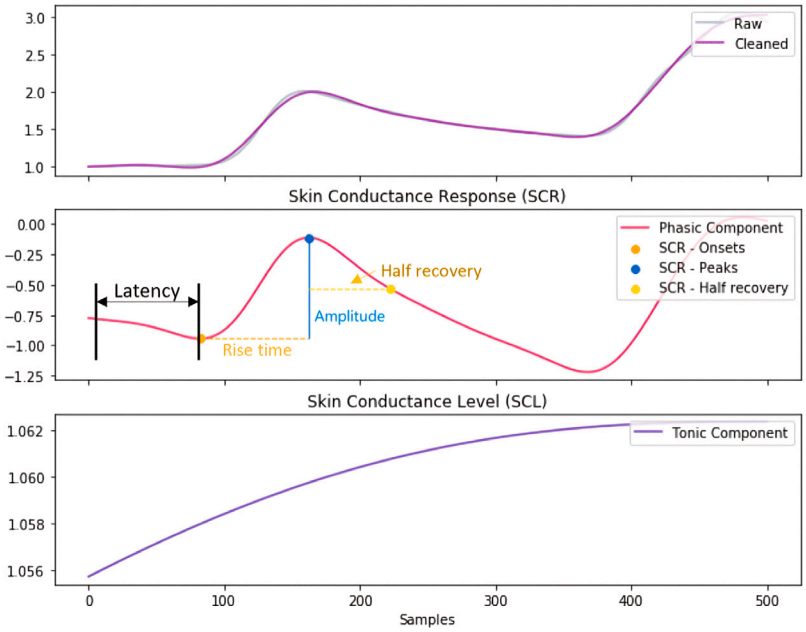

Fig. 2. An EDA signal can be broken down into two components and four characteristics.

(PPG), and electrodermal activity at 25, 500, 50, and 1 Hz intervals, respectively. VSM WaveTool, a program that runs on a personal computer, was used to record data, start and stop, and adjust other parameters of the logging. It is also possible to save synchronized multiparameter data on the internal memory of the ADI-VSM, and then retrieve it at a later time for offline analysis. The battery in the device has a capacity of 140 mAh, and its typical run time is 18 h with all of the sensors active. The battery may be recharged.

3.6. Benchmark datasets

Physiological signals from a variety of experiments are available in publicly available datasets. Signals are collected from several locations on the skin in these studies, as required by the devices and the questions being asked. Our stress detection method was built using three publicly available datasets: CLAS (Markova, Ganchev, & Kalinkov, 2019), VerBIO (Zhu et al., 2022), and WESAD (Schmidt, Reiss, Duerichen, Marberger, & Van Laerhoven, 2018), which contain EDA signals and a physiological questionnaire. Categorizing models are trained and tested using physiological signals included in the three datasets. Currently, EDA signals are used to analyze depression as an important factor (Ahn et al., 2019; Hong et al., 2010; Kim et al., 2020; Roy et al., 2012).

1. **CLAS:** Markova et al. (2019) Intelligent human–computer interaction (HCI) led to the development of the CLAS dataset. Emotion and stress detection are just two examples of the many automated human psychological and physiological assessments included in this data set.
2. **WESAD:** Schmidt et al. (2018) WESAD was developed to investigate whether or not it is possible to recognize emotional states based on physiological markers.
3. **VerBIO:** Zhu et al. (2022) The VerBIO data set was constructed with the intention of determining whether or not stress could have an effect on physiological signals present during public speaking.

The reasons for choosing these three datasets for comparison can be broken down into two main categories. (1) EDA data that are included in all three datasets. For VerBIO and WESAD, the Empatica E4 wristband was used to collect data. (2) These three datasets were used to predict emotions and stress. For these reasons, there were more opportunities to compare the data.

4. Methodology

Time spent gathering the data could thus vary from a few minutes to several hours, or possibly days. Given the high processing costs and uneven sample sizes, it is especially advantageous if the EDA data has a longer lifespan, which could make analysis easier. The pattern of pain experienced during labor is unique to each individual woman. Fig. 3 depicts the proposed structure from data acquisition to classification. The subject's EDA signals are gathered via a wrist-worn device, like a smartwatch. The data is transmitted wirelessly to an accessible computer or smartphone. The signal then passes through a series of signal processing steps, with the retrieved attributes ultimately being put to use in a classification process. The above process can be carried out within the framework using data prepossessing, artifact detection, data labeling, and classification using the novel ensemble model with neural networks to effectively predict prenatal depression.

- **Data Pre-processing**
  During the pre-processing stage of the data, there are primarily three phases that are carried out, as depicted in Fig. 3. This pre-processing is carried out using following steps such as Data segmentation, Components separation, and Attribute extraction.

  (1) **Data Segmentation:** In most cases, EDA data is gathered during various stages of the labor. Time spent gathering the data could thus vary from a few minutes to several hours, or possibly days. Given the high processing costs and uneven sample sizes, it is especially advantageous if EDA data has a longer lifespan, which could make analysis easier. As a consequence of this, EDA data have to be segmented to a particular length in order to maintain a consistent format for the samples and to lessen the amount of computing effort required. For the subsequent step of processing, this research divided all of the data and labels based on a non-overlapping sliding window of five seconds.
  (2) **Components Separation:** Further data processing is necessary after passing the continuously recorded EDA signal, raw-EDA, through a 5-Hz Butterworth low pass filter to remove high-frequency noise and redundant information. Typically, an EDA signal consists of two parts: (i) the SCL,

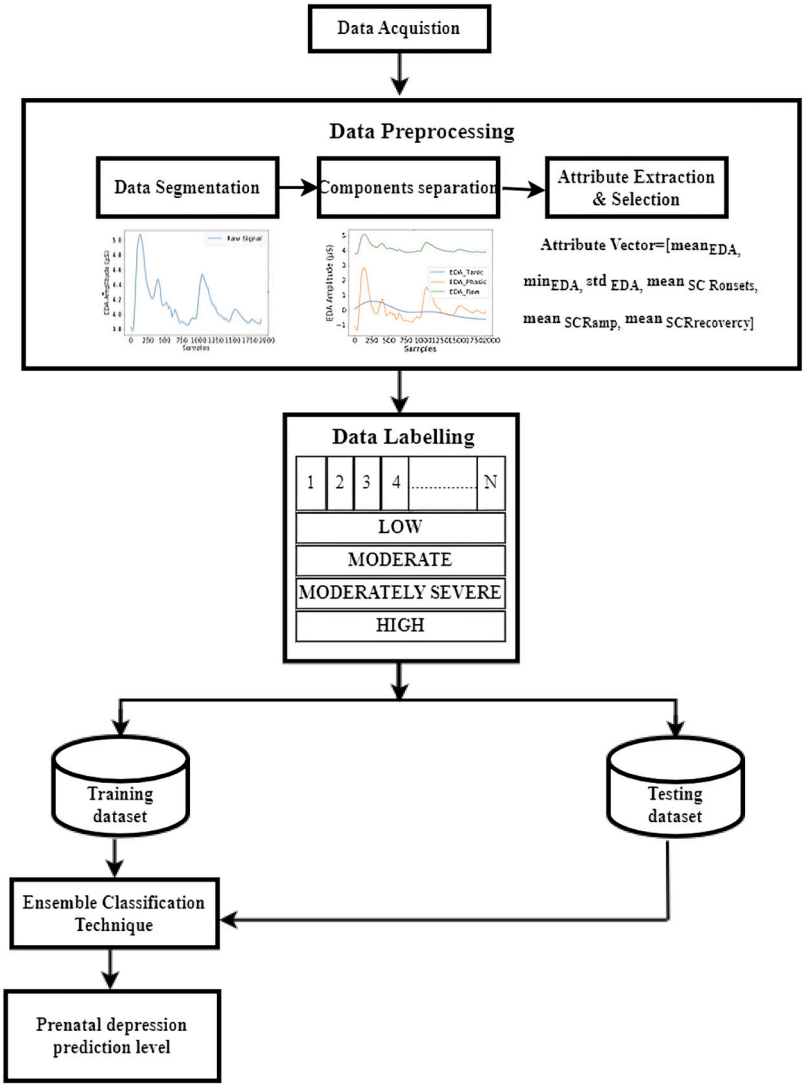

Fig. 3. The overall structure of the ensemble based deep learning model for depression detection.

or tonic EDA, which shows constant changes when no outside factors are present, and (ii) phasic EDA, which shows sudden changes when something is happening, either internally or externally. Skin conductance levels are the name given to both kinds of EDA. A phasic response known as the specific or event-related skin conductance response (SCR) happens in reaction to a singular and distinct external stimulus, like a startle event like a gunshot. The Event-Related Skin Conductance Response ER-SCR is another name for this reaction. Nonspecific phasic responses, also known as NS.SCRs, are responses that take place spontaneously and unprompted by any external stimuli, in contrast to SCRs. In human beings, the SCL and NS.SCR are taken into consideration as indicators of psychological activation. Therefore, in order to ensure accurate analysis in the future, it is necessary to first separate SCR and SCL components from the data and then apply artifact removal procedures. The cvxEDA model (Xia et al., 2015) is used to analyze the SCR and SCL parts. The foundations of this model are sparsity, convex optimization, and MAP. Additionally, the raw EDA signals are cleaned up of motion artifacts.

(3) **Attribute extraction & selection:** We calculated characteristics from each 5-s EDA window segment. Estimating statistical properties from the initial EDA and its first and second derivatives was the first stage. Two AR variables (n1 and n2) and the AR noise variance were used as features in a AR model that was employed to simulate the EDA sequence. When noise is introduced into EDA data, the amount of noise that remains in the AR model is larger compared to clean data, which is reason AR modeling is being used. This results in higher values for both AR parameters and free from motion artifacts. A high-resolution temporal frequency decomposition approach, VFCDM, was employed to enhance the dynamic aspects of both clean and damaged EDA (Wang, Siu, Ju, & Chon, 2006).
Using VFCDM for biosignal applications has shown useful in analyzing signal properties and reducing noise and artifacts (Hossain et al., 2021; Posada-Quintero, Florian, Orjuela-Cañon, & Chon, 2016). We separated EDA data segments into 12 distinct frequency bands that did not overlap using VFCDM. We used VFCDM to determine the two signals' ranges (max–min), as well as their mean, variance, and ratio of variances. In order to train the data, an attribute vector was constructed utilizing statistical attributes and additional SCR attributes. This was done because processing the signals with all of their properties would raise the computing cost. The data was subsequently trained using this attribute vector. According to Chen et al.

(2015) and Zhang et al. (2017), seven attributes are chosen for inclusion in the attribute vector. For example, the attribute vector can be expressed as:

$$\begin{aligned}\text{AttributeVector} = [&\text{mean}_{EDA}, \min_{EDA}, \max_{EDA}, \text{std}_{EDA},\\ &\text{mean}_{SC\ \text{Ronsets}}, \text{mean}_{SCRamp},\\ &\text{mean}_{SCR\ \text{recovery}}]\end{aligned} \tag{1}$$

where the actual EDA value in each signal frame is used to calculate the mean, minimum, maximum, and standard deviation of EDA (Chen et al., 2015).

Attribute selection was accomplished by the use of the RF machine learning method (Breiman, 2001). The RF method's interpretable, low over fitting, and excellent prediction accuracy make it a popular choice as an attribute selection algorithm. Embedded approaches, which include RF for attribute selection, are a hybrid of filter and wrapper methods. The embedded methods work well, are easy to generalize, and can be understood via the lens of attribute selection.

- **Data Labeling:**
  Data from each person's time-synchronized EDA and speedometer was manually sorted using a Matlab-based data visualization tool. A total of 125 observations were made using non-overlapping windows, and each one was categorized as either mild depression (mild), moderate depression (MOD), moderately severe (MOD-S), or high depression (HIGH). A comprehensive experimental evaluation that included five distinct window sizes (5 s, 10 s, 15 s, 30 s, and 60 s) led to the selection of the window as the optimal option. The MA sections were extracted from all of the signals, but the depression classes were only derived from the 10-min EDA data collected just before the two PHQ-9 surveys. When the following conditions were met during a 5-s epoch: (i) the skin conductivity level was zero or negative; (ii) the EDA signal showed a sudden maximum related to movement as indicated by accelerometer data; or (iii) the quantization error surpassed 5% of the signal amplitude, the epoch was designated as MA by Taylor et al. (2015).
  The depression components of the questionnaire were determined with the use of the scores that were received from the PHQ-9-Y1. The average score on the PHQ-9-Y1 for subject i throughout both surveys is represented by the notation SStextsubscriptij (where i can be any number from 1 to N and j can be either 1 or 2). The first poll used a j value of 1, whereas the second survey used a j value of 2. For this study, the total number of participants is represented by the letter M. On the PHQ-9-Y1, the available score range for each question extends from 5 to 27, with 27 being the highest possible score. In light of this, the formula SS ij = (SS ij-5)/27 was performed in order to calculate subject i's normalized depression indices. This was done in order to determine subject i's normalized depression indices. After the scores were calculated, the depression sections were categorized as 'LOW' (scores between 0.0 and 0.25) 'MOD' (scores between 0.26 and 0.50), 'MOD-S' (scores between 0.51 and 0.74), and 'HIGH' (scores between 0.75 and 1) according to the normative values given in the PHQ-9 handbook (Srisurapanont, Oon-Arom, Suradom, Luewan, & Kawilapat, 2023). Table 2 displays a summary of the gleaned characteristics.
  In the meantime, the data with all attributes is utilized to train the models and to compare the classification results with the extracted attribute vector, even if no attribute extraction is conducted on this data.
- **Training & Testing: subject-independent validation strategy**
  During the process of extracting attributes and ensemble model based classification, we used a validation technique that was independent of the subject (mothers). We used a method known as LOOCV for validation, which implies that for every fold, we withdrew one mother's data from testing and kept the others for training. We used a group M-fold validation to pick features. For each fold of the LOOCV validation, the training data was used for this validation. To ensure that the classifiers were completely subject-agnostic, we once again used a group M-fold. Using the group k-fold, we carried out a grid-search cross-validation approach in order to determine which parameter was the most suitable for each fold.
  In order to avoid the over fitting and test the model's efficacy on new data, the LOOCV validation approach is used while the model is being trained. The LOOCV validation method is said to be one of the most well-known and commonly utilized approaches. It is appropriate for use with very limited datasets and results in a model that is objective. Additionally, in comparison to other technologies, this one requires a comparatively shorter amount of time for calculation. The dataset is randomly split into M equal-sized pieces, D1, D2, D3, …, DM, using the LOOCV validation method. Afterwards, the model undergoes N rounds of training and testing, with a testing set consisting of one of the independent components. for some dataset where N is the number of samples. As seen in Fig. 4, the data is subsequently processed using an oversampling strategy. It is well-known that bootstrapping is one of the most used resampling processes, among numerous others that employ the replacement strategy to generate new samples or resamples from existing ones. By using the bootstrapping technique (Jain & Moreau, 1987), a total of 4900 records are obtained from the data instances of 189 moms. Next, the bootstrapped dataset is split into three sections: the training set (composing 70% of the total), the validation set (20%), and the test set (composing 10% of the total). A tenfold cross-validation strategy is then utilized. Additionally, the grid search approach has to be used in order to maximize the model's performance (Sun, Xue, Zhang, & Yen, 2018), other parameters, which are shown in Table 3, are also modified during each iteration. Once the training of the model has been completed successfully, the accuracy that was acquired in each fold is mathematically calculated using Eq. (3).

$$Acc_{cv}^{N} = \frac{1}{N} \sum_{(x_i, y_i) \epsilon F_i} \sigma\left(I\left(F_{S(i)}, x_i\right), y_i\right) \tag{2}$$

where $\sigma\left(I\left(F_{S(i)}, x_i\right), y_i\right)$ denotes the accuracy observed for each fold.

- **Methods for classifications**
  This section provides detail description of the proposed ensemble model and four different classification algorithms which is used to compare he effectiveness of the ensemble model for predicting the severity levels of the prenatal depression mothers.

  (A) **Artificial Neural Networks ANN**
  Designing a ANN or a MLP is inspired by the way information is processed in the human nervous system (Zhang, 2016). Each of these layers is classified into one of three groups: input, concealed, or output. While the output layer is in charge of assigning each input class to its corresponding input pattern, the input layer is in charge of defining that pattern. According to Karadeniz (2021), hidden layers are given weights to help fine-tune the network and decrease the amount of mistake.

  (B) **K-Nearest Neighbor (KNN):**
  By evaluating the distance between the values of characteristics, KNN can categorize data. Specifically, we want to find out, from a training dataset, which K training examples closely resemble the input instance. We will use the majority distribution of the preceding K instances to

determine the categorization of the input case. Our cross-validated grid-search method determined that 3 is the optimal K value (Machhale, Nandpuru, Kapur, & Kosta, 2015). The model can be scored with a variety of parameters, and the best-performing parameter will be chosen as the final one by the technique.

(C) **Decision Tree**

The decision tree is an example of supervised learning, which is a type of learning that may be used to problems involving classification as well as regression. Although decision trees are straightforward and show promise for managing high-dimensional data, they are quite unstable. Even a little change to the data may have a big impact on the overall structure. Another drawback of the decision tree is the extensive training time it requires. Sing (2015) made a decision. When building the decision tree, entropy and information gain are used as measures for selecting attributes. For data categorization purposes, at each level, the attribute with the lowest entropy is picked. When the entropy of a branch drops to zero, it is called a leaf node; otherwise, it will keep branching out. Mathematically, entropy for different qualities may be calculated using the following formula:

$$E = -\sum_{i=1}^{n} p_i \times \log_2 p_i \tag{3}$$

(D) **Random Forest**

Additional supervised machine learning algorithms include the random forest approach. For this method, the decision tree is the backbone of the forest construction process. Several separate decision trees are combined into one ensemble to form the random forest method (Liaw, Wiener, et al., 2002). This is where random forest comes in handy; it uses replacement to train each tree on a random sample, so it can avoid decision trees' instability and data sensitivity. This method is sometimes called bagging. Decision trees and random forests differ in another respect: the attributes they provide are not uniformly random. As it builds its hierarchical structure, the decision tree considers all features. The random forest, in contrast, trains each tree using a subset of really random features.

(E) **Stacked Ensemble based Deep Learning (EBDL)**

The ensemble learning method is a combination of many machine learning approaches that uses the predictions of several base models to enhance predictive performance (Wang, Hao, Ma, & Jiang, 2011). The groundwork model might have been built using any of the available machine learning methods. Homogeneous ensemble learning models are those that are trained utilizing identical basis learners. However, an ensemble is considered heterogeneous if its base learners vary from one another. There are three different kinds of algorithms that make up ensemble learning: bagging, boosting, and stacking. Stacking approach involves training the basic classifiers on the same dataset and then using an extra classifier called a meta-learner to boost the model's performance. A single-level stacking strategy is used in the present inquiry, and several deep learning models, referred to as the Ensemble based Deep Learning model, are used in the first phases of the process. In the end, the predictions for the presence or absence of prenatal depression are provided by a Logit-Boost that has been fitted using the predictions of the separate classification models, as explained in Algorithm 1. The stacking ensemble seen in 3 is constructed using deep neural networks. As stacking based models provides model diversity (Barton & Lennox, 2022), Flexibility (Lu et al., 2023), interpretability (Alarfaj & Khan, 2023) among other models. There are 14 nodes in the input layer and 4 nodes in the output layer of the artificial neural network (ANN) being built for the proposed research. Using the intra-user variability as a basis, this ANN aims to predict the severity levels. By using an activation function called ReLU and three hidden layers, a neural network may be built as shown in Table 3

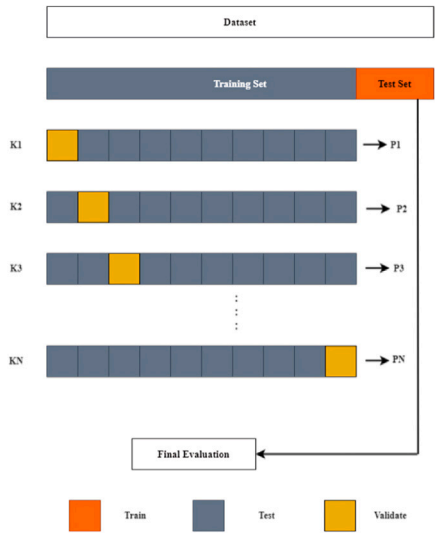

Fig. 4. Subject-independent validation strategy.

Algorithm 1 Ensemble based Deep learning (EBDL) model's Algorithm

Require: Training dataset D, where $D = \{D_1, D_2, D_3 \ldots D_m\}$
Ensure: Prediction of stress level from the stacking ensemble classifier

1: Divide D into M equal halves at random in a way that $D = D_1, D_2, D_3, ..., D_M$
2: **for** $m = 1$ to M **do**
3: Make use of D to train base classifiers, and then repeat steps 4–7.
4: Determine the weighted total and include bias into every hidden layer node by
Info $= \sum_i^n x_i \times W_i + \text{bias}$
5: Calculate the values of $\Delta W = W - \eta \frac{\partial E}{\partial W}$ and $E = \frac{1}{2}\sum_{p=1}^{n}\sum_{o=1}^{m}\left(T_{io} - A_{io}\right)^2$
6: To obtain the lowest mistake rate, tweak the values of the learning parameters and weights.
7: Apply a ReLU activation function f(Info)= max(0, Info) at each base classifier.
8: **end for**
9: Formulating the training set for meta-classifier.
10: **for** $t = 1toT$ **do**
11: $D_E = \{\mathbf{x}i', y_i\}$, where $\mathbf{x}_i' = \{h_{k1}(\mathbf{x}i), h_{k2}(\mathbf{x}i), \ldots, h_{kT}(\mathbf{x}_i)\}$
12: **end for**
13: Develop LogistBoost, a meta-learning classifier, by using D_E
14: Return Predictions $y_i = \{y_1, y_2, y_3 \ldots y_n\}$ from the derived EBDL model

In addition, the depression prediction scheme associated with that group is involved. Consequently, the proposed system classifies the incoming test instances into relevant subgroups and then activates the related classifier to forecast the amount of depression. Fig. 3 shows the hierarchical structure of the multilevel depression detection method, which begins with an artifact detection phase and then uses a

Table 2
Dataset statistics.

Variable		Train dataset	Test dataset	Total dataset
No of subjects		100	89	189
5s Epochs	LOW	4646	2721	7367
	MOD	4321	3256	7577
	MOD-S	4047	3217	7264
	HIGH	4756	4117	8873
	Total	17 770	13 311	31 081
% Epochs		76%	24%	100%
Age (years)		34.4 ± 19.8	34.2 ± 19.7	34.3 ± 19.5
Height (cm)		154.9 ± 8.2	156.2 ± 7.4	155.55 ± 7.8
Weight (kg)		66.6 ± 8.2	62.7 ± 9.7	64.65 ± 8.9

depression classification framework to partition data. This framework is based on the suggestion that the multilevel depression detection method should be used. There will be motion artifacts, because the EDA sensors will shift somewhat on the skin due to both the body's motion and the skin's moisture.

4.1. Evaluation metrics

One of the most important factors that determines the effectiveness of an ensemble model is its ability to produce correct results. In order to determine how well the proposed model works, a wide variety of performance assessment metrics are used.

- Precision: Precision is defined as the degree to which measurements are in close proximity to each other.

$$Precision = \frac{TruePositive}{(TurePositive + FalsePositive)} \tag{4}$$

- Recall:The ability of the model to identify true positives for each of the given classes is measured by the recall mechanism.

$$Recall = \frac{TruePositive}{(TurePositive + FalseNegative)} \tag{5}$$

- Specificity: The capability of the model to determine the actual negatives associated with each possible class is measured by specificity.

$$Specificity = \frac{TrueNegative}{(TureNegative + FalsePositive)} \tag{6}$$

- Accuracy: Accuracy takes into account the frequency with which the proposed machine learning model properly classifies an instance of data that has not yet been observed.

$$Accuracy = \frac{TruePositive + TrueNegative}{(TruePositive + TureNegative + FalsePositive + FalseNegative)} \tag{7}$$

- RMSE:The difference between the actual values and the anticipated values is what the RMSE evaluates.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \hat{x}_i\right)^2} \tag{8}$$

- MAE:It is the absolute variation between the actual values and the expected values that is measured by the mean absolute error.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|x_i - \hat{x}_i\right| \tag{9}$$

The predicted and actual values are denoted by xi and xi, respectively, in this context, where N is the number of occurrences. In the previously given equations, the letters TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Prenatal depression of a mother can be accurately predicted by looking at the number of true positives and true negatives. By contrast, the sum of the ensemble model's false positive and false negative rates is what ultimately decides how many erroneous predictions it produces.

Table 3
Hyperparameters.

Parameter	Values
Input layer	14
Output layer	4
Hidden layer	6
Activation function	ReLU
Learning rate	0.01
Optimizer	Adam
#epochs	100

5. Results

This section describes the experimental step-up along with the results obtained from the model to predict prenatal depression based on the EDA signals from the wearable wrist device. It is subdivided into six subsections as follows:

5.1. Experimental set up

The experimental implementation and performance analysis of the suggested model are the topics that are presented in this Simulation Setup. Experiments are carried out on a system as follows: an Intel(R) Core(TM) i7-9050H processor, a primary memory capacity of 8 GB, a clock frequency of 2.60 GHz, an NVIDIA GeForce GTX 1050 GPU, and a 64 bit Windows-10 operating system. The proposed model is implemented using various application programming interfaces (API) that are available in the most recent version of Python, which is 3.9.

5.2. Statistical data analysis

The activation of the Hypothalamic-Pituitary-Adrenal (HPA) axis that occurs when an individual is exposed to a stressor causes the body to begin producing cortisol. Salivary cortisol has emerged as a valid biomarker of sympathetic activation during times of depression, as a result of research conducted in recent years. To determine the levels of cortisol in the individuals' systems prior to delivery, repeated samples of their saliva were taken during the delivery period. The SOMA Bioscience salivary cortisol test kit (Anusha et al., 2019) was used for both the collection of samples and their subsequent measurement

At a high level, this section summarizes the data that will be considered in the following. See Fig. 5 for box plots of the moms going through labor pains and the cortisol levels of 189 people. A common trend during the active stage of labor is a gradual increase in cortisol levels, which signals a rising stress level. The only way salivary cortisol was used to support the claim that respondents felt more stressed was as an objective measure. The high degree of subject-to-subject variability prevented its use for purposes such as generating a more nuanced range of stress levels.

Cortisol concentrations at different stages of work are displayed as box plots for 189 patients. T bars show the spread of the data for each box plot. Within the box is the median, while the vertical line denotes the interquartile range. The center is shown by the red dot.

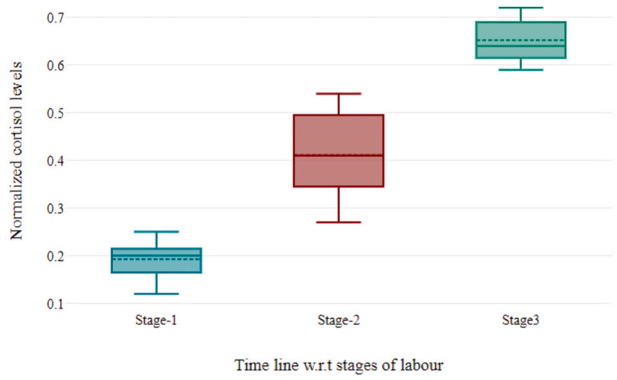

Fig. 5. Box plot of cortisol level varying.

Table 4
Ensemble models' relative performances.

Algorithm	Precision	Recall	Specificity	Accuracy	RMSE	MAE
Decision tree	0.7765	0.7736	0.8369	0.7482	0.52	0.36
KNN	0.7436	0.7234	0.8126	0.7319	0.50	0.34
Random forest	0.8752	0.8154	0.8791	0.8159	0.44	0.32
ANN	0.9367	0.8596	0.9052	0.8876	0.41	0.28
EBDL model	**0.9578**	**0.9254**	**0.9523**	**0.9387**	**0.31**	**0.24**

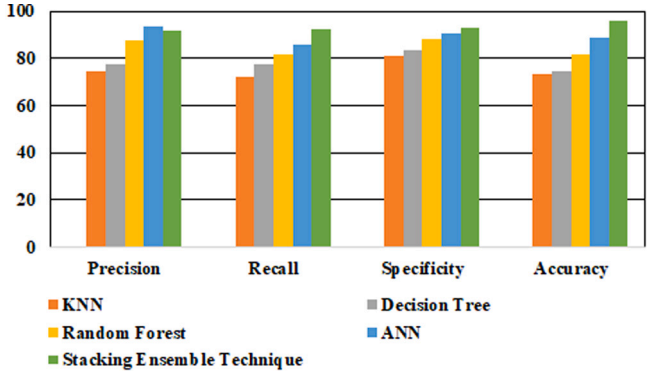

Fig. 6. Comparative analysis of ensemble model.

5.3. Comparative analysis of stacking based EBDL model with others classification algorithms

In order to prove that our work on prenatal depression prediction is worthwhile, we will compare how well the suggested EBDL model performs against the baseline machine learning techniques. To begin, a LOOCV validation strategy is utilized for LOOCV to perform training and validation of all baseline machine learning and ANN models. The baseline machine learning algorithms include DT, RF, KNN, and ANN.As mentioned in Subsection named data labeling, we compare the results of these algorithms' testing with our suggested ensemble model, which is derived from a range of performance assessment measures. Table 4 contains the stated results of the calculations.

Moreover, Fig. 6 shows the findings of a bar graph comparing the different performance metrics achieved by the strategies that were considered.

5.4. Evaluation of accuracy of ensemble based deep learning model on benchmark datasets

Tables 5, 6, 7 display the accuracy of the ensemble-based Deep Learning model on benchmark datasets such as CLAS, WESAD and VerBIO respectively. Importantly, the Ensemble-based Deep Learning model predicts stress accurately in the CLAS, VerBIO, and WESAD datasets when compared to previous researches. Using EDA from the collected dataset, the Ensemble-based deep learning model maintains its optimal overall result at 93.87%. It appears that EDA has the potential to mirror the emotional changes that occur in mothers during delivery. Furthermore, unlike SCR, variations in ECG and PPG might not be as sensitive to minor mood changes. Therefore, when it comes to emotion-related detection, EDA should be the first choice with the Ensemble-based Deep Learning model.

Table 5
Detection accuracy for EDA modality of CLAS benchmark dataset.

Method	Accuracy	F1 score	AUC
KNN (Radhika & Oruganti, 2021)	0.699	0.7026	0.6959
ANN (Radhika & Oruganti, 2021)	0.7261	0.7434	0.7321
RNN (Radhika, Subramanian, & Oruganti, 2022)	0.889	0.8262	0.7914
CRNN (Radhika et al., 2022)	0.8925	0.7831	0.7216
EBDL	0.912	0.935	0.914

Table 6
Detection accuracy for EDA modality of WESAD benchmark dataset.

Method	Accuracy	F1 score	AUC
KNN (Zhu et al., 2022)	0.664	0.412	–
ANN (Zhu et al., 2022)	0.857	0.785	–
RNN (Radhika et al., 2022)	0.8652	0.8309	0.7982
CRNN (Bobade & Vani, 2020a)	0.8432	0.8071	0.7654
EBDL	0.8962	0.8234	0.8126

Table 7
Detection accuracy for EDA modality of VerBIO benchmark dataset.

Method	Accuracy	F1 score	AUC
KNN (Zhu et al., 2022)	0.5882	0.5421	0.5268
DCNN (Zhu et al., 2022)	0.6175	0.5987	0.5628
RNN (Zhu et al., 2022)	0.8011	0.7967	0.8074
CRNN (Zhu et al., 2022)	0.8643	0.8071	0.8106
EBDL	0.8957	0.8724	0.8214

5.5. Evaluation based on ablation concepts

The use of ablation principles was done to guarantee the significance of the innovations of stacking Ensemble-based deep learning procedures in this model. The novelties of this model were: (1) removal of motion artifacts. (2) Include depression levels based on the severity score of the women undergo delivery of the child. (3) Ensemble-based deep learning model based on both subject dependent & subject independent validation strategy. For this purpose, the above three strategies are combined into one offered, which then produces a variety of different combinations and is executed using the collected datasets and the accuracy, accuracy, recall and F1 score were evaluated as presented in Table 8. These combinations include: (1) without artifact removal + without LOOCV (2) artifacts removal + without LOOCV (3) artifacts removal + with LOOCV. It is very clear from the Table 8 that artifacts removal with LOOCV subject dependent & subject independent validation strategy provides better results with the proposed ensemble-based Deep Learning model that provides the best accuracy of all baseline classifiers.

6. Discussion

The discussion section is divided into four different subsections based on the evaluation results obtained by executing the Ensemble-based Deep Learning model with both subject dependent & subject independent validation strategy with the collected data set to predict prenatal depression in women. This section elaborates the necessity of the proposed method, which helps to solve existing problems such

Table 8
Evaluation based on Ablation concepts.

Methods	Criteria	Random Forest	ANN	EBDL
Without artifacts removal+ without LOOCV	F1 score	57.12	52.42	52.17
	Precision	51.68	47.15	52.35
	Recall	42.67	42.36	49.77
	Accuracy	52.37	50.84	48.62
Without Artifacts removal+ with LOOCV	F1 score	52.74	51.62	58.12
	Precision	56.74	49.82	52.68
	Recall	48.61	47.96	51.76
	Accuracy	56.47	59.96	**59.21**
Artifacts removal+ without LOOCV	F1 score	52.74	51.62	58.12
	Precision	53.74	52.36	60.24
	Recall	42.61	57.42	59.01
	Accuracy	58.47	62.89	**62.37**
artifacts removal+ LOOCV	F1 score	81.64	87.28	92.19
	Precision	86.48	82.51	91.54
	Recall	82.34	80.94	94.62
	Accuracy	81.59	88.76	**93.87**

as removal of motion artifacts and generality of the subject. It also provides better prediction with deep learning models.

6.1. Interpretation of ensemble based deep learning model with collected vs. standard three benchmark datasets

Compared to ECG, PPG, and signal combinations, EDA provides a more accurate prediction of depression before surgery (Anusha et al., 2019). As a result, we used EDA as a useful signal to predict depression levels in women during labor. One EDA signal serves as the foundation for all other benchmark datasets, which provides the best performance with the ensemble-based deep learning model compared to baseline learning models. The datasets collected to analyze prenatal depression produce better classification results in all classifiers. One possible interpretation of this finding is that due to the efficient motion artifacts removal model as well as the hybrid subject dependent & subject independent validation strategy. Moreover, when the different classifiers are assembled, it produces better results than traditional baseline models. It also provides the severity level of prenatal depression in women rather than without including severity levels it classifies as binary classification models. However, if the severity levels segregated by the PDSS score from the collected data set provides better accurate classification levels with prenatal depressed mothers.

6.2. Interpretation based on ensemble based deep learning classification model

Think about it: Compared to other baseline algorithms, the Ensemble-based Deep Learning stacking method that was suggested does the best job and has the highest predicted accuracy of 93.79%. It should be noted that the suggested technique outperforms DT, RF, SVM, and ANN based on the findings of the F1 score (92. 19%), precision (91. 54%) and recall (94. 62%). In addition, two statistical approaches are used, namely the RMSE and MAE, to compare the results. You may calculate the mean absolute error using any of these approaches. With RMSE (0.31) and MAE (0.24) values that are lower than any of the baseline algorithms, the suggested technique stands out. After comparing the two sets of data, one may conclude that the stacking Ensemble-based Deep Learning used in this work outperforms the baseline models in every way.

6.3. Interpretation based on ablation concepts

For the purpose of ensuring that the innovative approaches in this model are given the importance they deserve, ablation concepts were utilized. (1) The elimination of motion artifacts was one of the notable innovations of this model. (2) Include the levels of depression according to the severity of the women who are going through the process of giving birth. (3). Ensemble-based deep learning model with neural networks based on both subject dependent & subject independent validation strategy. From Table 8, it becomes evident that the removal of artifacts, incorporating severity levels, and the Ensemble-based Deep Learning model based on both subject dependent & subject independent validation strategy yields the highest level of accuracy among all classifiers.

From the Table 8 without artifact removal, thought, and without a combined validation model the accuracy results are not more than 50%, which means that without cleaned EDA signals it is very difficult to predict vital clues about prenatal depression effectively. In view of the combination of the inclusion of artifacts removal, and without subject dependent & subject independent validation strategy provides a slightly better prediction since the analysis is carried out with cleaned EDA signals. The artifacts removal, including subject dependent & subject independent validation strategy does not provide classification as binary classification (Taylor et al., 2015) each and every mother has their own specific characteristics such as sweat gland density and skin thickness, so all these factors contribute much to the prediction of prenatal depression.

6.4. Advantages & disadvantages

This proposed method contains the following metrics:

- As it predicts the degrees of severity of depression, this study functions similarly to multiclassification.
- It is also possible to predict different mood disorders using this EBDL model.
- This model can be considered for horizontal deployment to reconstruct the diagnostic process of other mental disorders (such as major depressive disorder, schizophrenia, bipolar disorder, and dementia) provided that suitable attribute extraction methods are employed.

Some notable drawbacks of this study are listed as follows.

- In this study, we focus on EDA to detect changes in maternal prenatal depression. Active continuous monitoring is needed to make this prediction accurately.
- Limitation of a wrist-worn wearable device in terms of accessible modalities and assessment of the possible contribution of various psychological signals to stress detection.
- Furthermore, as spectrum information in the frequency domain might reflect oscillation information, frequency-domain characteristics may provide higher discriminating ability of the psychological responses than time-domain features for nonstationary psychological signals.

6.5. Future work

The same categorization algorithms can also be used with the various modalities included in a wearable wrist computer. As a result, we were able to evaluate how various psychological signals may have contributed to the identification of stress. As most smartwatches have these sensors, we focus on ECG, PPG, and EDA to identify emotional changes in the wearer. Adopting Explainable AI (XAI) techniques to interpret our model prediction is our future direction, which could increase the trust of the results produced by deep-leaning models.

7. Conclusion

The results of this research imply that prenatal depression is predicted using a stacking ensemble learning model based on deep neural networks for the early prediction of prenatal depression. It is implemented during childbirth using the wrist EDA. The two-stage approach, which involves artifact identification first and then a subject-dependent and independent validation strategy along with the ensemble model, was able to successfully classify 93.79 percent of the new dataset. Several baseline machine learning techniques are used to evaluate the suggested model's performance. One way to assess performance is through evaluation metrics. Common metrics include recall, accuracy, specificity, and precision. Beyond that, two statistical metrics are tested, the MAE and the RMSE, are tested. The suggested stacking ensemble method predicts depression severity levels with a accuracy rate of 93. 79%. The results for precision (91. 54%), F1 score (92. 19%), and recall (94. 62%), compared to baseline learning methods, demonstrate that the suggested method outperforms them. Furthermore, the suggested model is robust, as shown by the minimal values of RMSE (0.31) and MAE (0.24). Therefore, this model can be used to forecast the severity of prenatal depression based on EDA signals. These preliminary results are intriguing because they suggest that EDA will be more helpful than previously thought for stress monitoring in actual settings. Intriguingly, this study found that the association between EDA and depression was higher among first-time mothers than among mothers who had previously given birth. Other case studies, including daily activity tracking, senior care, fitness assistance, and telemonitoring programs, can be implemented around this prenatal mental health monitoring.

CRediT authorship contribution statement

Abinaya Gopalakrishnan: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Xujuan Zhou:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Revathi Venkataraman:** Writing – review & editing, Validation, Supervision. **Raj Gururajan:** Writing – review & editing, Validation, Supervision. **Ka Ching Chan:** Writing – review & editing, Validation, Supervision. **Niall Higgins:** Writing – review & editing, Validation, Supervision, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ahn, J. W., Ku, Y., & Kim, H. C. (2019). A novel wearable EEG and ECG recording system for stress assessment. *Sensors*, *19*(9), 1991.

Alarfaj, F. K., & Khan, J. A. (2023). Deep dive into fake news detection: Feature-centric classification with ensemble and deep learning methods. *Algorithms*, *16*(11), 507.

Anusha, A., Sukumaran, P., Sarveswaran, V., Shyam, A., Akl, T., Preejith, S., et al. (2019). Electrodermal activity based pre-surgery stress detection using a wrist wearable. *IEEE Journal of Biomedical and Health Informatics*, *24*(1), 92–100.

Ayers, S. (2004). Delivery as a traumatic event: prevalence, risk factors, and treatment for postnatal posttraumatic stress disorder. *Clinical Obstetrics and Gynecology*, *47*(3), 552–567.

Barton, M., & Lennox, B. (2022). Model stacking to improve prediction and variable importance robustness for soft sensor development. *Digital Chemical Engineering*, *3*, Article 100034.

Beck, C. (1998). A checklist to identify women at risk for developing postpartum depression. *Journal of Obstetric, Gynecologic, and Neonatal Nursing*, *27*, 39–46.

Benfield, R. D., Newton, E. R., Tanner, C. J., & Heitkemper, M. M. (2014). Cortisol as a biomarker of stress in term human labor: physiological and methodological issues. *Biological Research for Nursing*, *16*(1), 64–71.

Bobade, P., & Vani, M. (2020a). Stress detection with machine learning and deep learning using multimodal physiological data. In *2020 second international conference on inventive research in computing applications* (pp. 51–57). IEEE.

Bogomolov, A., Lepri, B., Ferron, M., Pianesi, F., & Pentland, A. (2014). Pervasive stress recognition for sustainable living. In *2014 IEEE international conference on pervasive computing and communication workshops (PERCOM WORKSHOPS)* (pp. 345–350).

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Broeders, J. H. (2017). Transition from wearable to medical devices. In *Analog devices*. Inc.

Caparros-Gonzalez, R. A., Romero-Gonzalez, B., Strivens-Vilchez, H., Gonzalez-Perez, R., Martinez-Augustin, O., & Peralta-Ramirez, M. I. (2017). Hair cortisol levels, psychological stress and psychopathological symptoms as predictors of postpartum depression. *PLoS One*, *12*(8), Article e0182817.

Chen, W., Jaques, N., Taylor, S., Sano, A., Fedor, S., & Picard, R. (2015). Wavelet-based motion artifact removal for electrodermal activity. In *2015 37th annual international conference of the IEEE engineering in medicine and biology society* (pp. 6223–6226).

Federenko, I., & Wadhwa, P. (2004). Women's mental health during pregnancy influences fetal and infant developmental and health outcomes. *CNS Spectrums*, *9*(3), 198–206.

Gao, S., & Lima, D. (2022). A review of the application of deep learning in the detection of alzheimer's disease. *International Journal of Cognitive Computing in Engineering*, *3*, 1–8.

Gopalakrishnan, A., Gururajan, R., Venkataraman, R., Zhou, X., Ching, K., Saravanan, A., et al. (2023). Attribute selection hybrid network model for risk factors analysis of postpartum depression using social media. *Brain Informatics*, *10*(1), 28.

Gopalakrishnan, A., Venkataraman, R., Gururajan, R., Zhou, X., & Genrich, R. (2022). Mobile phone enabled mental health monitoring to enhance diagnosis for severity assessment of behaviours: a review. *PeerJ Computer Science*, *8*, Article e1042.

Gopalakrishnan, A., Venkataraman, R., Gururajan, R., Zhou, X., & Zhu, G. (2022). Predicting women with postpartum depression symptoms using machine learning techniques. *Mathematics*, *10*(23), 4570.

Grace, S. L., Evindar, A., & Stewart, D. (2003). The effect of postpartum depression on child cognitive development and behavior: a review and critical analysis of the literature. *Archiv Women Mental Health*, *6*, 263–274.

He, C., Levis, B., Riehm, K., Saadat, N., Levis, A., Azar, M., et al. (2020). The accuracy of the patient health questionnaire-9 algorithm for screening to detect major depression: an individual participant data meta-analysis. *Psychotherapy and Psychosomatics*, *89*(1), 25–37.

Hernandez, J., Riobo, I., Rozga, A., Abowd, G., & Picard, R. (2014). Using electrodermal activity to recognize ease of engagement in children during social interactions. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing* (pp. 307–317).

Hobel, C., Goldstein, A., & Barrett, E. (2008). Psychosocial stress and pregnancy outcome. *Clinical Obstetrics and Gynecology*, *51*(2), 333–348.

Hong, S., Yang, Y., Lee, J., Yang, H., Park, K., Lee, S., et al. (2010). Ambulatory stress monitoring with a wearable bluetooth electrocardiographic device. *Studies in Health Technology and Informatics*, *161*, 66–76.

Hossain, M. B., Bashar, S., Lazaro, J., Reljin, N., Noh, Y., & Chon, K. (2021). A robust ECG denoising technique using variable frequency complex demodulation. *Computer Methods and Programs in Biomedicine*, *200*, Article 105856.

Huangfu, B., & Cheng, W. (2025). Cognitive computing method based on decoding psychological emotional states. *International Journal of Cognitive Computing in Engineering*, *6*, 32–43.

Jain, A., & Moreau, J. (1987). Bootstrap technique in cluster analysis. *Pattern Recognition*, *20*(5), 547–568.

Kim, J., Park, J., & Park, J. (2020). Development of a statistical model to classify driving stress levels using galvanic skin responses. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *30*(5), 321–328.

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R News*, *2*(3), 18–22.

Lu, M., Hou, Q., Qin, S., Zhou, L., Hua, D., Wang, X., et al. (2023). A stacking ensemble model of various machine learning models for daily runoff forecasting. *Water*, *15*(7), 1265.

Ma, Y., Liu, Q., & Yang, L. (2024). Machine learning-based multimodal fusion recognition of passenger ship seafarers' workload: A case study of a real navigation experiment. *Ocean Engineering*, *300*, Article 117346.

Machhale, K., Nandpuru, H., Kapur, V., & Kosta, L. (2015). MRI brain cancer classification using hybrid classifier (SVM-knn). In *2015 international conference on industrial instrumentation and control* (pp. 60–65).

Markova, V., Ganchev, T., & Kalinkov, K. (2019). Clas: A database for cognitive load, affect and stress recognition. In *2019 international conference on biomedical innovations and applications* (pp. 1–4).

Masoudi, Z., Kasraeian, M., & Akbarzadeh, M. (2022). Assessment of educational intervention and acupressure during labor on the mother's anxiety level and arterial oxygen pressure of the umbilical cord of infants (PO2). a randomized controlled clinical trial. *Journal of Education and Health Promotion*, *11*.

McMahon, C. A., Barnett, B., Kowalenko, N. M., & Tennant, C. C. (2006). Maternal attachment state of mind moderates the impact of postnatal depression on infant attachment. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *47*(7), 660–669.

Miller, N., Asali, A. A., Agassi-Zaitler, M., Neumark, E., Eisenberg, M. M., Hadi, E., et al. (2019). Physiological and psychological stress responses to labor and delivery as expressed by salivary cortisol: a prospective study. *American Journal of Obstetrics and Gynecology*, *221*(4), 351–e1.

Naegelin, M., Weibel, R., Kerr, J., Schinazi, V., La Marca, R., Wangenheim, F., et al. (2023). An interpretable machine learning approach to multimodal stress detection in a simulated office environment. *Journal of Biomedical Informatics*, *139*, Article 104299.

Nielsen, D., Videbech, P., Hedegaard, M., Dalby, J., & Secher, N. (2000). Postpartum depression: Identification of women at risk. *BJOG Interbnational Journal of Obstet Gynaecol*, *107*, 1210–1217.

Oubrahim, Z., Amirat, Y., Benbouzid, M., & Ouassaid, M. (2023). Power quality disturbances characterization using signal processing and pattern recognition techniques: A comprehensive review. *Energies*, *16*(6), 2685.

Poh, M. Z., Swenson, N., & Picard, R. (2010). A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *IEEE Transactions on Biomedical Engineering*, *57*(5), 1243–1252.

Posada-Quintero, H., Florian, J., Orjuela-Cañon, & Chon, K. (2016). Highly sensitive index of sympathetic activity based on time-frequency spectral analysis of electrodermal activity. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *311*(3), R582–R591.

Radhika, K., & Oruganti, V. (2021). Deep multimodal fusion for subject-independent stress detection. In *2021 11th international conference on cloud computing, data science & engineering (confluence)* (pp. 105–109).

Radhika, K., Subramanian, R., & Oruganti, V. (2022). Joint modality features in frequency domain for stress detection. *IEEE Access*, *10*, 57201–57211.

Roberts, S. L., Roberts, S. L., Bushnell, J. A., Roberts, S. L., Bushnell, J. A., Collings, S. C., et al. (2006). Psychological health of men with partners who have post-partum depression. *Australian & New Zealand Journal of Psychiatry*, *40*(8), 704–711.

Roy, J. C., Boucsein, W., Fowles, D., & Gruzelier, J. (2012). *Progress in electrodermal research*: *vol. 249*, Springer Science & Business Media.

Rykov, Y. G., Patterson, M. D., Gangwar, B. A., Jabar, S. B., Leonardo, J., Ng, K. P., et al. (2024). Predicting cognitive scores from wearable-based digital physiological features using machine learning: data from a clinical trial in mild cognitive impairment. *BMC Medicine*, *22*(1), 36.

Salkind, N. (2010). *Encyclopedia of research design*: *vol. 1*, sage.

Sandman, C. A., Glynn, L. M., & Davis, E. P. (2016). Neurobehavioral consequences of fetal exposure to gestational stress. Fetal development: Research on brain and behavior, environmental influences, and emerging technologies. (pp. 229–265).

Saylam, B., & Incel, Ö. D. (2024). Multitask learning for mental health: Depression, anxiety, stress (DAS) using wearables. *Diagnostics*, *14*(5), 501.

Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., & Van Laerhoven, K. (2018). Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction* (pp. 400–408).

Srisurapanont, M., Oon-Arom, A., Suradom, C., Luewan, S., & Kawilapat, S. (2023). Convergent validity of the edinburgh postnatal depression scale and the patient health questionnaire (PHQ-9) in pregnant and postpartum women: Their construct correlations with functional disability. In *Healthcare* (p. 699).

Sun, Y., Xue, B., Zhang, M., & Yen, G. (2018). An experimental study on hyper-parameter optimization for stacked auto-encoders. In *2018 IEEE congress on evolutionary computation* (pp. 1–8).

Taylor, S., Jaques, N., Chen, W., Fedor, S., Sano, A., & Picard, R. (2015). Automatic identification of artifacts in electrodermal activity data. In *2015 37th annual international conference of the IEEE engineering in medicine and biology society* (pp. 1934–1937).

Wang, G., Hao, J., Ma, J., & Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, *38*(1), 223–230.

Wang, H., Siu, K., Ju, K., & Chon, K. (2006). A high resolution approach to estimating time-frequency spectra and their amplitudes. *Annals of Biomedical Engineering*, *34*, 326–338.

Xia, V., Jaques, N., Taylor, S., Fedor, S., & Picard, R. (2015). Active learning for electrodermal activity classification. In *2015 ieee signal processing in medicine and biology symposium (spmb)* (pp. 1–6).

Zhang, Z. (2016). A gentle introduction to artificial neural networks. *Annals of Translational Medicine*, *4*(19).

Zhang, Y., Haghdan, M., & Xu, K. (2017). Unsupervised motion artifact detection in wrist-measured electrodermal activity data. In *Proceedings of the 2017 ACM international symposium on wearable computers* (pp. 54–57).

Zhu, L., Ng, P., Yu, Y., Wang, Y., Spachos, P., Hatzinakos, D., et al. (2022). Feasibility study of stress detection with machine learning through eda from wearable devices. In *ICC 2022-IEEE international conference on communications* (pp. 4800–4805).

Zhu, L., Spachos, P., Ng, P., Yu, Y., Wang, Y., Plataniotis, K., et al. (2023). Stress detection through wrist-based electrodermal activity monitoring and machine learning. *IEEE Journal of Biomedical and Health Informatics*.